Anirudh Cheruvu

anirudhcheruvu2014@gmail.com +1-667-379-9852 Linkedin Github Raleigh, NC

Education

University of North Carolina at Charlotte — Charlotte, North Carolina
M.S. in Computer Science GPA: 3.9/4.0 — *Aug 2022 - May 2024*

Jawaharlal Nehru Technological University — Hyderabad, India
B.Tech. in Computer Science GPA: 8.27/10.0 — *Aug 2018 - July 2022*

Skills

AI & Machine Learning: NLP, LLMs, RAG, Hugging Face Transformers, Prompt engineering, LangChain, PyTorch, TensorFlow, scikit-learn, FAISS

Backend & APIs: Python, Flask, FastAPI, REST, Node.js; Testing with PyTest (TDD); Git

Cloud, Data & DevOps: AWS (S3, Lambda, Step Functions, SQS, CloudWatch); CI/CD (Jenkins); Docker; DBs: Oracle, MySQL, DynamoDB

Projects

GraphifyAI - Agentic AI-Powered Data Visualization Tool Demo — GitHub

- Developed an **agentic AI system** (Python, **FastAPI**, **Gemini Live API**) for real-time data analytics & visualization, processing voice queries (speech recognition/NLP), performing autonomous database queries (**API tool-calling**), and delivering interactive visualizations with Generative AI speech/vision.
- Engineered real-time visualization of on-screen tables via a **multimodal LLM's OCR** capabilities for structured data extraction and dynamic visualization generation.
- Designed a **Streamlit web interface** with Chrome extension APIs (for audio input/screen capture) and **WebSockets** for real-time AI backend communication.

CineGraphAI: LLM-Powered GraphRAG Movie Knowledge Chatbot GitHub

- Developed an LLM-powered chatbot with an interactive **Streamlit** front-end, using **GraphRAG** over a **Neo4j** graph of IMDb movies to answer complex natural language queries.
- Leveraged **Google Gemini API** to parse queries and autonomously generate **Cypher** to query **Neo4j**, reasoning over relationships among movie entities to produce context-aware responses.

The Small-Model Reasoning Study GitHub

- Benchmarked the `Qwen3-0.6B` (650M) model on the `GSM8K` dataset, comparing prompting styles (direct, chain-of-thought, few-shot, and a chat template with "think" tags) and training approaches (full supervised fine-tuning vs. LoRA) using scriptable pipelines.
- Results: accuracy improved from 8% (direct) to 54% with chain-of-thought; the "think" template after full fine-tuning reached 66.7%. Full SFT outperformed LoRA variants (48% vs. 45%). Released code, checkpoints, and evaluation artifacts.

Experience

Trensyc LLC — United States
AI developer — *Apr 2025 - Present*

- Currently driving POCs for innovative AI applications in generative AI, agentic systems, and advanced retrieval augmented generation (RAG), using LangChain, Huggingface Transformers and LLMs.

Radiance Technologies LLC — United States
Data Engineer - Trainee — *Aug 2024 - Apr 2025*

- Trained in the AWS SA – Associate learning path with exposure across compute, storage, integration, workflow, messaging, and monitoring.
- Built and maintained SQL queries and Python (Flask) API endpoints to power internal services. Implemented request validation and unit tests, set up CloudWatch alerts, and worked through Git/Jenkins CI/CD in an Agile team.

Cognizant Technology Solutions — Hyderabad, India
Web Development Intern — *Feb 2022 - Jun 2022*

- Completed **400-hour** Full-Stack Web Development training (Cognizant Early Engagement), mastering the **MERN** stack through **5+** web applications and creating reusable, performant **React** UI components.